Social Network Analysis and Mining (2019) 9:30
https://doi.org/10.1007/s13278-019-0576-8

ORIGINAL ARTICLE

Characterising and evaluating dynamic online communities from live microblogging user interactions

Hugo Hromic[1] · **Conor Hayes**[1]

Received: 22 December 2018 / Revised: 17 June 2019 / Accepted: 19 June 2019 / Published online: 3 July 2019

Abstract
Microblogging social media focuses on fast open real-time communication using short messages between users and their followers. These platforms generate large amounts of content, and community finding techniques are a suitable alternative for organising it. However, there is no clear agreement in the literature for a definition of *user community* for the microblogging use case, leading to unreliable ground-truth data and evaluation. In this work, we differentiate between *functional* and *structural* definitions of communities for microblogging. A functional community groups its users by a common independent social function, e.g. fans of the same football team, while in a structural community the members exclusively depend on their connectivity in a network, e.g. modularity. We build and characterise eight types of functional communities to be used as user-labelled ground-truth and five types of user interactions networks from Twitter. We then evaluate—in static and dynamic scenarios—thirteen popular structural community definitions using five different Twitter datasets, exploring their goodness and robustness for detecting the functional ground-truth under different perturbation strategies. Our results show that definitions based on internal connectivity, e.g. Triangle Participation Ratio, Fraction Over Median Degree or Conductance work best for the Twitter use case and are very robust. On the other hand, other scores such as Modularity are limited and do not perform well due to the sparsity and noise of microblogging. Furthermore, using user activity as basis to separate communities into active *hotspots* further improves the performance of community detection in microblogging.

1 Introduction

Online Social Networks (OSN) have developed from simple static blogs into richer interactive systems such as Facebook, Twitter, LinkedIn or YouTube, to name a few. Microblogging is a type of OSN which allows for fast real-time open broadcasting of short content among friends and/or millions of followers. Examples are Twitter, Weibo (the Chinese counterpart of Twitter) and Tumblr, which is similar to Twitter but focused on multimedia posts. Twitter is currently one of the most widely known microblogging OSN in the world, with more than 330 million monthly active users as of December 2017 (Aslam 2018), generating an average of 500 million *Tweets* (short messages) per day.

This massive amount of content fuels an increasing disorganisation for its users that can be alleviated with content filtering or clustering techniques such as community detection (Palla et al. 2005; Papadopoulos et al. 2011). However, a fundamental challenge is the diversity in the literature for a definition of *user community*, which makes community detection difficult to evaluate and interpret. For microblogging OSN, researchers often use the same definition as for more traditional social media (Darmon et al. 2015; Bakillah et al. 2015; Amor et al. 2016; Cao et al. 2015), or definitions from topic analysis and user profiling (Zhou et al. 2012; Akbari and Chua 2017). However, we argue that these adoptions might not be suitable for microblogging without considering the temporal dynamics of the platform due to its particular fast pace and user sparsity.

Instead of attempting to craft yet another community definition specifically for microblogging, we will prefer and evaluate a more flexible non-personal wider interpretation. Our hypothesis is that in microblogging, people do not seek to be closely related but instead are more curious about the collective opinion of the masses. Therefore, we adopt the differentiation between *functional* and *structural* definitions

✉ Hugo Hromic
hugo.hromic@insight-centre.org

Conor Hayes
conor.hayes@insight-centre.org

[1] Insight Centre for Data Analytics, NUI Galway, Galway, Ireland

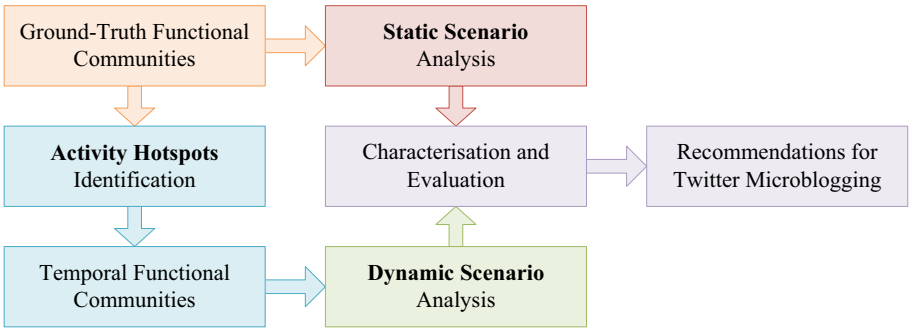


Fig. 1 Proposed workflow for the research carried in this work. Static and dynamic scenarios are both considered

of user communities found in Yang and Leskovec 2015. A functional community groups its users by a common and independent social *function*, e.g. fans of the same football team, while in a structural community, the members exclusively depend on their *connectivity in a network*, e.g. their average node degree. The task of community detection is then defined for our purposes as the discovery of functional communities from structural patterns in a network of user interactions.

In this work, we propose to characterise and investigate the network structure of functional communities considered as user-labelled ground-truth for the microblogging OSN in both, *static* and *dynamic* scenarios. The communities under study can emerge from real-time microblogging user interactions such as replies, user mentioning, posts rebroadcasting and quoting.

Static structural sources, e.g. follower networks, are commonly used for this purpose; however, they are often prohibitive to capture for global analysis and are not suitable for detecting fast-paced community formation and termination (Darmon et al. 2015). Therefore, we propose to use live streams of interactions instead. Furthermore, we continue our previous work in (Hromic and Hayes 2018) by incorporating the temporal characteristics of microblogging in our investigation. In particular, we identify user activity *hotspots* in the ground-truth communities that later are used to improve their detection using traditional structural scoring functions.

A general diagram of the proposed research methodology in this work can be found in Fig. 1. First, a set of ground-truth functional communities is defined and constructed for the case of microblogging using real-world Twitter social streams. Then, a baseline characterisation and evaluation of these communities is performed on a static scenario, i.e. without considering any temporal aspects. Afterwards, a methodology for identifying user activity hotspots in the ground-truth is proposed and a new set of temporal functional communities is assembled. A second evaluation is performed with these temporal communities under a dynamic scenario that now considers the time dimension. Finally, the two scenarios under study are contrasted and a set of recommendations for community detection in Twitter microblogging social networks is offered.

The main motivation of this work is that it is difficult to identify user communities in live streams of microblogging OSN due to their velocity and low reciprocal characteristics. We hypothesise that the highly dynamic and fast-paced nature of microblogging causes users to switch or lose interest about topics quickly, rendering conventional community discovery based on more static and dense networks less effective. Research work and applications for microblogging data that rely on community detection approaches (Gupta et al. 2012; Hromic et al. 2017) must account for the dynamics of microblogging for better results. Therefore, we propose the following research questions.

(**RQ1**) How can independent, explicitly user-labelled, ground-truth functional communities and networks of live user interactions be captured from Twitter?

(**RQ2**) Do the defined ground-truth functional communities in (RQ1) evidence distinctive structural patterns in the associated user interactions networks?

(**RQ3**) How well-existing state-of-the-art structural community definitions (e.g. based on triangle participation, conductance or modularity) align to the defined ground-truth functional communities in (RQ1), including their robustness to random perturbations?

(**RQ4**) Using the dynamic user activity in time as a basis, how can activity *hotspots* be identified in the defined ground-truth functional communities in (RQ1) to be used for identifying time-scoped sub-communities?

(**RQ5**) How well the same structural community definitions in (RQ3) align to the newly identified sub-communities based on activity hotspots in comparison with the original ground-truth in (RQ1)?

Addressing these research questions provides a number of contributions to the research field of community detection for microblogging, including dynamic community detection,

visualisation and evaluation. The identified contributions are described below.

1. From (RQ1), a methodology for building ground-truth functional communities from microblogging live user interactions, particularly from Twitter.
2. From (RQ2), (RQ3) and (RQ5), an in-depth characterisation and evaluation of global and structural properties for functional communities in microblogging, for both the static and the dynamic scenarios.
3. From (RQ3) and (RQ5), a set of recommendations on community detection algorithms based on data-driven evaluation of real-world Twitter data.
4. From (RQ4), a strategy for the identification of temporal activity hotspots in user communities in microblogging based on the user interactions network that improves the performance of existing structural community detection approaches designed for more dense and static data, without modifying them.
5. An open-source implementation of the analytical framework developed available on request.

The results presented in this work are based on representative examples of captured datasets according to the experiment under discussion. However, the complete set of results can be found in Hromic (2019).

The rest of this paper is organised as follows: In Sect. 2, we present background and related work; then, in Sect. 3, we establish and examine sets of ground-truth functional communities for Twitter, including their definitions. In Sect. 4, we study the performance of a range of existing structural community definitions over our ground-truth and propose recommendations for the microblogging case in a static scenario. In Sect. 5, we investigate the dynamic scenario using activity hotspots in microblogging communities. Finally, we present our conclusions and propose future directions in Sect. 6.

2 Background and related work

2.1 Characteristics of microblogging social media

OSN can be categorised according to the degree of social functionalities they offer (Kietzmann et al. 2011): identity, conversations, sharing, presence, relationships, reputation and groups. For example, Facebook relies on strong friendship relationships, Foursquare focuses on physical user presence, YouTube focuses on video content sharing and LinkedIn values user identity.

In the case of Twitter, relationships between users are less personal, promoting instead a more open ambient for socialising (Shamma et al. 2009). Twitter users can post short messages publicly and other users can reply, quote and *retweet* (rebroadcast) them. There is a limit of 280 characters per post, prompting users for brevity and clarity in their content.

A user can choose to follow another user for the up-to-date content, often with no approval needed or without requiring to be followed back; however, this mechanism has been found to be of low reciprocity (Kwak et al. 2010). For example, only 20% of users follow each other in contrast to other services such as Flickr (70%) or Yahoo! 360 (80%), which show much more reciprocal connections (Cha et al. 2009; Kumar et al. 2010). Information in Twitter spreads less than five hops away, shorter than in other known OSN (Kwak et al. 2010), highlighting the strength of microblogging as a medium for rapid information diffusion compared to other OSN focused on verified relationships. All the above suggests that Twitter Followers Networks might not be adequate for traditional structural community detection that rely on more static and dense networks.

Microblogging also differentiates itself from characteristics of classic human social networks (Newman and Park 2003): the distribution of subscribers is not power law, the degree of separation is shorter, and most links between its users are not reciprocated (Kwak et al. 2010). However, Twitter still evidences degrees of homophily, i.e. contact between similar people occurs at a higher rate than among dissimilar members (McPherson et al. 2001), resembling communities.

2.2 User communities in microblogging

In OSN, users can develop natural groupings by finding other users with similar interests (Tang and Liu 2010), i.e. the principle of homophily. Moreover, communities allow these users to better focus on interesting content.

The definition of a user community for Twitter is broadly described in the literature as *"a group of nodes that are more densely connected to each other than to nodes outside of the group"* (Tang and Liu 2010; Java et al. 2007; Gupta et al. 2012; Lu and Brelsford 2014; Sakaki et al. 2010; Shamma et al. 2009; Yang and Leskovec 2015). However, communities can emerge from very diverse motives and intentions, and are often based on topical subjects or shared interests, i.e. they are functional to the users (Java et al. 2007; Yang and Leskovec 2015). Furthermore, these functional communities have been also suggested to require an intermediate social object that connects people together to truly become social, otherwise they lose interest (Engeström 2005).

The community detection task for microblogging is mostly addressed by means of exploiting static networks, e.g. followers, captured in snapshots (Java et al. 2007; Gupta et al. 2012; Lu and Brelsford 2014; Shamma et al. 2009; Sakaki et al. 2010). In this work, instead we aim to

understand how user communities can form solely through their public live user interactions represented as a network. Community detection has been also approached via a combination of both methods (Darmon et al. 2015; Bakillah et al. 2015; Amor et al. 2016); however, we argue that such static networks are expensive to retrieve and maintain fresh in comparison to a stream of messages (Darmon et al. 2015).

User role understanding is a long-standing field in social network analysis. For example, the interpretation of user communities using ego networks (Mcauley and Leskovec 2014) and the identification of user roles in communities (Amor et al. 2016) have been previously studied. In these works, functional classes of users are defined based on their structural position in the network. In this work, we interpret this concept of social functionality around social objects that can emerge in microblogging.

2.3 Dynamic communities in microblogging

Processing microblogging social networks often involves large-scale and highly dynamic data that should not be disregarded (Sundaram et al. 2012; Lu and Brelsford 2014; Amor et al. 2016; Aslak et al. 2018). Sundaram et al. provide a comprehensive review of these dynamic characteristics using scalable methods. They observe that real-world communities are based on coherent and sustained interactions in time and identify the need for dynamic temporal methods. Similarly, Lu and Brelsford later investigate the dynamics of microblogging in response to extreme events such as natural disasters. Their findings reveal that the community joining and leaving behaviours are not random effects. Users tend to stay in their current state and are less prone to shift communities or join new ones when in solitude, and user communities quickly shift their topics under the effect of the disaster event. In this work, we leverage this behaviour by identifying activity hotspots that community detection approaches can use for mining.

The study of dynamic user roles in microblogging has been also proposed (Amor et al. 2016), as well as the adoption of multilayer social networks (Aslak et al. 2018). Amor et al. propose graph-based methods to process Twitter data captured in the context of political debates around controversial public matters. The authors propose a method that considers directionality and are able to discover robust communities that correlate with the observed information flow. In this work, instead we use undirected networks to represent live user interactions to also obtain robust user communities. For the case of multilayer networks, Aslak et al. present an approach for the InfoMap method (Rosvall and Bergstrom 2008) that estimates layer dependencies at the network layer level. When constraining the network information using the above estimates, InfoMap is able to better uncover user communities with nodes in multiple groups. Their findings suggest that attaching network layers at the node-level can improve the modelling of information diffusion in highly dynamic scenarios such as microblogging, which improves the discovery of intermittent structures. This approach relies on mining friendship networks, our work instead focuses on the dynamics of user interactions.

The work in Myers and Leskovec 2014 further examines the dynamics of Twitter. The authors proposed a model for the dynamics of the network structure using steady rates of change that are disturbed by sudden bursts based on information diffusion. Their model can predict which events of information diffusion will produce bursts in the underlying network. More generally, dynamics have been also studied in other types of social and non-social networks (Palla et al. 2007).

2.4 Study methodology

We inspire our research methodology on the works of Palla et al. 2007; Yang and Leskovec 2015. The authors empirically study structural community definitions on classical OSN in both, static and dynamic scenarios. However, our work differentiates from theirs in that: (1) we extend the original study from traditional OSN to the microblogging case, taking into account the particularities and dynamics of the platform, and (2) we adapt the original experiments to address the challenges imposed by microblogging, including data volume and its different characteristics.

3 Ground-truth communities in Twitter

In this Section, we address research question (RQ1): how can independent ground-truth functional communities and networks of live users be captured from Twitter? We distinguish two independent definitions of user communities: *functional*, based on social function, and *structural*, based on the connectivity in a network.

Functional communities will represent our ground-truth data because users themselves explicitly state the social function of their posts, e.g. referencing the same hashtag or mentioning the same celebrity. For this to be meaningful, any given social function must be repeated at least a certain number of times. On the other hand, a structural community is a set of users with a particular connectivity pattern in the underlying live interactions network, e.g. a high edge density or triangle count.

Our goal is to investigate the relationship between these two definitions of communities, considering the task of community detection as the recovery of user communities based on a structural definition that later correspond to ground-truth functional communities (Yang and Leskovec 2015). In other

words, we aim to find an alignment of connectivity patterns in the interactions network to explicitly labelled social functions.

When using real-world data as ground-truth for community detection evaluation, i.e. as proposed in this work with microblogging functional community definitions (detailed in Sect. 3.2), care must be taken as this paradigm can be potentially problematic (Peel et al. 2017). In particular, a poorly performing community detection algorithm might not be fully explained by the ground-truth data as there might be other factors affecting its performance. For example, certain social functions in microblogging might be irrelevant to the underlying group structure or the ground-truth might capture social aspects that are not purely structural. Therefore, in this work, we also aim to investigate how the constructed microblogging ground-truth is modelled by structural community definitions based on different families (detailed in Sect. 4.2).

3.1 Building live interactions networks

In Twitter, posts can be composed using special syntax. Users can provide searchable *#hashtags*, mention other users using *@username* anchors, link to web resources and embed media files, e.g. pictures or videos. Some of this special syntax, together with replying to posts and retweeting, can be used to form a network of interactions between users (Hromic and Hayes 2014).

Based on Hromic et al. 2015 and Yang et al. 2014, we consider four concrete types of Twitter interactions for building networks of live interactions from a stream of Tweets: mentions, quotes, replies and retweets. We note that these interactions are not used as social indicators nor ground-truth in the network, but simply as social ties between users to form structures. For simplicity, we will consider the interactions networks as undirected.

Therefore, an undirected network $G = (V, E, W)$ is proposed with a set of user vertices $u \in V$, interaction edges $e = (u_i, u_j) \in E$ and temporal edge weights $w(e, t) \in W$. The complete procedure for building this network of interactions from a stream of Tweets is presented in Algorithm 1. Every time t, a user u_i interacts with another user u_j using any of the defined interaction types, an edge $e = (u_i, u_j)$ is added to the network and the edge weight $w(e, t)$ is incremented by one. To reduce the amount of space required for storing these potentially numerous temporal weights, a quantisation function $Q(t, q)$ is applied to each time observation. In other words, weights W account user activity per minute, hour, day or any quantisation q required.

Algorithm 1: Construction of a weighted social user interactions network with quantised time-aware and typed edges from a stream of Tweets.

Data: a stream of Tweets $\mathbb{T} = \{tw_0, tw_1, \ldots, tw_n\}$
Data: a time quantisation q in seconds, e.g. 3600
Result: an interactions network $G = (V, E, W)$ for $\mathbb{T}$
$V \leftarrow \emptyset$, $E \leftarrow \emptyset$, $W \leftarrow \emptyset$
Function `addToNetwork`$(u_i, u_j, time, type)$ **is**
  $e \leftarrow (u_i, u_j)$
  $V \leftarrow V \cup \{u_i, u_j\}$, $E \leftarrow E \cup \{e\}$
  if $w(e, t, type) = \emptyset$ **then**
    $w(e, t, type) \leftarrow 0$ `/* initialise weight if edge not yet seen */`
  end
  $w(e, t, type) \leftarrow w(e, t, type) + 1$
end
`/* The mentions, replyTo, retweetOf and quoteOf functions extract all the mentions, any replied user, any retweeted user and any quoted user from a given Tweet respectively. */`
for $tw_i \in \mathbb{T}$ **do**
  $author \leftarrow$ `author`(tw_i)
  $time \leftarrow$ `Q(time`(tw_i, q)`)` `/* quantise the time of posting of the Tweet */`
  for $m_i \in$ `mentions`$(tw_i) -$ `replyTo`$(tw_i) -$ `retweetOf`$(tw_i) -$ `quoteOf`(tw_i) **do**
    `addToNetwork`$(author, m_i, time, \{mentions\})$
  end
  if `replyTo`$(tw_i) \neq \emptyset$ **then**
    `addToNetwork`$(author,$ `replyTo`$(tw_i), time, \{replies\})$
  end
  if `retweetOf`$(tw_i) \neq \emptyset$ **then**
    `addToNetwork`$(author,$ `retweetOf`$(tw_i), time, \{retweets\})$
  end
  if `quoteOf`$(tw_i) \neq \emptyset$ **then**
    `addToNetwork`$(author,$ `quoteOf`$(tw_i), time, \{quotes\})$
  end
end

Initially, we considered building separate networks for each interaction type. However, a pair-wise network overlap analysis in our experimental data revealed an average low overlap of $\approx 3.82\%$, mostly between the *mention* and *reply* interaction types.

3.2 Building ground-truth functional communities

Ground-truth functional communities are built from a stream of Tweets where the members explicitly use a common functional social object of a particular type, independent of their underlying interactions. For example, if a set of users $\{u_1, u_2, u_3\}$ use the same common hashtag h, then a ground-truth community $C_h = \{u_1, u_2, u_3\}$ is

constructed. Later in Sect. 3.3, when using user activity hotspots, this behaviour is implicitly required to sustain a certain number of times for the ground-truth communities to stay relevant.

The following social objects are considered for building ground-truth functional communities from Twitter:

- *Hashtags* to group users that use the same hashtag, e.g. a trending topic or event of interest.
- *Mentions* to group users that mention the same user, e.g. a celebrity in a recent event.
- *Retweets* to group users that retweet the same user, e.g. a newscaster posting controversial news.
- *Quotes* to group users that quote the same user, e.g. provide an opinion over a third-party statement.
- *Countries*, *Cities* and *Places* to group users posting from the same location at different granularities. In Twitter, a *place* is an optional location entity that can be embedded in Tweets. Places can contain country and city attributes that can be used to form groups based on these abstractions.
- *URLs* to group users that share the same web link, e.g. an interesting cooking recipe.

Users might not be fully aware of these social objects creating connections between them in the form of functional communities. We investigate if such connection really exists through their live interactions.

Even though the mentions, retweets and quotes social functions are also used to build the interactions network and could be considered as an inherent bias, we emphasise that in contrast to the interactions network construction, instead here the mentioned, retweeted and quoted users themselves are used as social objects for members of the ground-truth communities to get together, similar to how they would be associated via a common topic modelled through a hashtag. For example, consider a ground-truth functional community $C_m = \{u_1, u_2, u_3\}$ of users mentioning the same user u_m. In this case, u_m does not need to be in this community nor interact with its members in any way. Instead, u_m is considered as an external motive for members of C_m to be connected socially.

Furthermore, we impose two basic build restrictions for the functional communities: (1) each group must have at least three members to facilitate the study of community scoring functions based on triad participation, and (2) communities with more than one connected component in the underlying live interactions network are separated and treated independently.

3.3 Activity hotspots in communities

The time dimension is an important aspect when dealing with microblogging social data as it is known to be fast paced and sparse. Therefore, we also investigate the dynamic scenario of microblogging in Sect. 5. Given a long enough period of time, network data captured from microblogging platforms tend to become complex and disorganised. In turn, mining user communities from these convoluted networks becomes more difficult. Therefore, we hypothesise that by finding user activity *hotspots* in these networks, it is possible to better focus on periods during the lifetime of user communities where it is easier to discover them.

As introduced in Sect. 3.1, during the capture of live user interactions networks $G = (V, E, W)$ from Twitter, we discretise and record the user activity using the weighting function $w(e, Q(t, q)) \in W$, for edges $e = (u_i, u_j) \in E$ between users $u_i, u_j \in V$. A quantisation parameter q is used to discretise the observation times t, e.g. by the minute, hour or day.

To investigate user activity hotspots, we first extract the user activity from the user interactions networks, based on the ground-truth functional communities. In preliminary experiments, we observed that this raw user activity is not suitable for temporal hotspots analysis due to noise and requires smoothing. Therefore, we adopt an exponential decay smoothing function that activates during user activity and decays in time during inactive periods (Palla et al. 2007). The smoothing function is introduced below.

$$w_{u,v}(t) = \sum_i w_i \exp(-\lambda |t - t_i| / w_i) \tag{1}$$

In Eq. 1, the summation runs over all the interactions recorded between users u and v and w_i is the weight of each interaction i observed at time t_i. The constant λ is a decay time characteristic.

Because it is not practical to empirically choose a λ time characteristic for each ground-truth functional community C, we estimate it based on the user activity of the community to obtain a fitted exponential curve. For this, we first add the weights $w(e, t)$ for each t of all the interactions related to C into a single $w'(e, t), e \in C$. Afterwards, we iterate over all pairs of adjacent observations $w'(e, t_i), w'(e, t_{i+1})$ computing the average absolute weight difference over the difference of their observed times t, as shown in Eq. 2.

$$\lambda(C) = \sum_i \frac{|w'(e, t_i) - w'(e, t_{i+1})|}{t_i - t_{i+1}} \tag{2}$$

3.4 Experimental ground-truth datasets

In this work, we investigate real-world Twitter data streams under different settings and periods of time. The

Table 1 Summary of built Twitter ground-truth datasets

Dataset	Timespan	Nodes	Edges	Communities	A_u	A_i	A_c
Pope2013	≈ 2 days	238,368	303,742	11,580	2.1082	0.6604	18.7042
Pope2013-Spl	≈ 2 days	6,593,649	6,140,684	40,812	4.0370	0.5098	27.1706
WorldCup2014	≈ 34 days	6,932,106	15,854,811	361,559	114.1077	32.3110	334.9153
RTE2015	≈ 63 days	643,292	1,446,852	56,025	163.6430	84.8706	687.7683
Ireland2017	≈ 245 days	1,067,982	2,826,754	62,562	1483.4306	355.7120	3881.9658

A_u is the average active user time, A_i is the average interaction time and A_c is the average community time. All times are in hours. Microblogging is noted for short-lived interactions

Twitter Streaming API offers two modes for collection: the *filter* and the *sample* endpoints. The first can retrieve streams using defined keywords, geographical coordinates and users to follow, while the latter provides a small random sample of all public Tweets being published in the platform. Data captured using this method have been compared to the global activity of Twitter as a whole (Morstatter et al. 2013). The findings suggest that the Streaming API can be biased depending on the amount of specified listening terms. However, it can also be accurate if the terms are specific enough to the topics or events of interest.

In total, we collected five streams from Twitter to be studied in this work. Using the filter endpoint, we captured two streams for two major worldwide events (wide-audience data), one stream based on geo-location (topic independent data) and a fourth stream for different seasonal TV shows and their audience (temporal periodicity data). To complement our study, we also captured one additional stream from the sample endpoint. Our experimental datasets are detailed below.

1. *Pope Conclave* (Pope2013)
   Captured during the Catholic Pope Conclave event in 2013. Spans for ≈ 2 days and contains 460K Tweets and 285K users. The filter endpoint was used and configured to listen for event-related hashtags and users to follow such as the hashtags #Conclave, #HabemusPapam, and the Twitter accounts of newscasters and candidate cardinals.
2. *Pope Conclave Sample* (Pope2013-Spl)
   Captured in parallel to Pope2013 using the sample endpoint. Spans for the same days and contains 9.9M Tweets and 8.8M users.
3. *FIFA World Cup* (WorldCup2014)
   Captured during the FIFA World Cup event in 2014. Spans for ≈ 34 days but contains 27.1M Tweets and 8M users. The filter endpoint was used and configured to listen for event-related hashtags and users to follow such as the hashtags #WorldCup, #Brazil2014, and the Twitter accounts of newscasters and participating football teams.
4. *RTÉ TV Programmes* (RTE2015)
   RTÉ is the public TV and Radio broadcaster of Ireland. We captured Tweets related to different TV programmes being broadcasted live by RTÉ. Spans for ≈ 63 days and contains 2M Tweets and 720K users. The filter endpoint was used and configured to listen for manually curated hashtags, keywords and users to follow such as #GreysAnatomy and #ThewalkingDead, related to each TV programme. Moreover, the filtering terms were dynamically configured according to the broadcasting time of each TV programme.
5. *Ireland* (Ireland2017)
   Captured using the location filter set for Ireland during 2017. Spans for ≈ 245 days and contains 7.7M Tweet and 1M users. It was captured using the geo-box $(-10.6696, 51.4199, -5.9947, 55.4351)$.

A summary of the network properties and number of built ground-truth communities can be found in Table 1. A total of 6,164,356 ground-truth communities were built from Twitter. In all the datasets, we chose a quantisation $q = 3600$ for recording time observations aligned in *hours* time unit. Note that the average user activity and interaction times are very short for the Pope2013 and Pope2013-Spl datasets.

3.5 Properties of ground-truth datasets

Before evaluating our constructed ground-truth functional communities from Twitter, we present a characterisation of their basic properties, with the purpose of providing a better understanding of the ground-truth.

In particular, we are interested on the sizes, memberships, overlap sizes and ages distributions for our ground-truth functional communities. The complementary cumulative distribution function (CCDF) in Eq. 3 computed for all of these properties in all of our Twitter datasets can be seen in Fig. 2. The CCDF represents the chances that a random variable X, e.g. community size, is valued above a range of values of interest x, e.g. communities larger than 1, 10, 100, etc.

$$CCDF(X) = P(X > x), X \text{ a random variable} \quad (3)$$

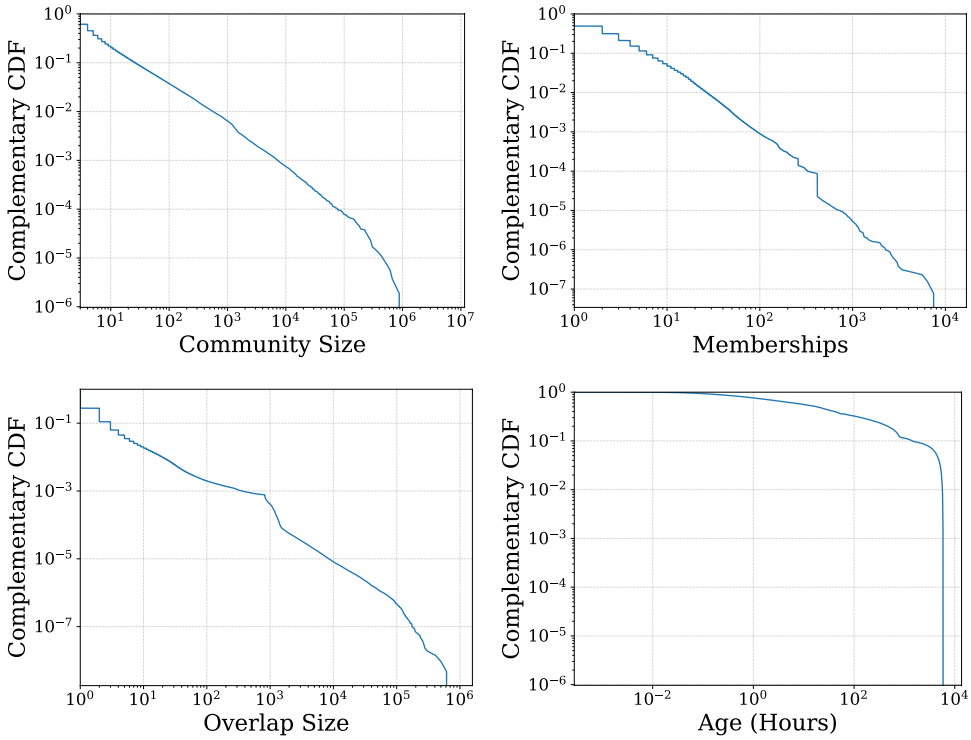


Fig. 2 Complementary Cumulative Distribution Function for the combined community sizes, membership sizes, overlap sizes and community ages across all of our Twitter datasets

In our results, all the CCDF distributions are skewed with most ground-truth communities being small, however larger communities also exist, e.g. sizes $> 10^4$. User memberships are sparse, with many users belonging to just a handful of communities and few users belonging to many communities. Regarding absolute user overlap sizes, i.e. the number of users in overlaps, again we observed a skewed distribution, following a power law, similar to Palla et al. 2005 for *detected* communities in contrast to *ground-truth* data.

We also investigate the relative size of user overlaps. This global property is useful because it can characterise how the ground-truth functional communities actually overlap: in nested structures or only for a small number of users (Yang and Leskovec 2015). For this, we measure the fraction of the size of the overlap between any two communities over the size of the smaller of these communities, $f = |C_i \cap C_j|/min(|C_i|, |C_j|)$. If $f \approx 0$, then the majority of the communities do not overlap, and if $f \approx 1$, then the communities have a nested structure where the smaller communities are incorporated into the larger groups. The results can be seen in Fig. 3, where a histogram of the fraction of overlap f for all of our ground-truth data is presented.

In this analysis, our ground-truth functional communities from Twitter are revealed as mostly not overlapping. This is expected as many of these communities are built around specific social objects. Nevertheless, a measurable number of ground-truth communities still exhibit some degree of nested overlapping $f \in [0.5, 0.7]$, evidencing groups of users that participate in Twitter using multiple social objects at the same time.

4 Evaluating communities: static scenario

In this Section, research questions (RQ2) and (RQ3) are addressed: do structural patterns exist in a network of interactions from Twitter that align to ground-truth functional communities? and how do existing structural community definitions perform in these networks?

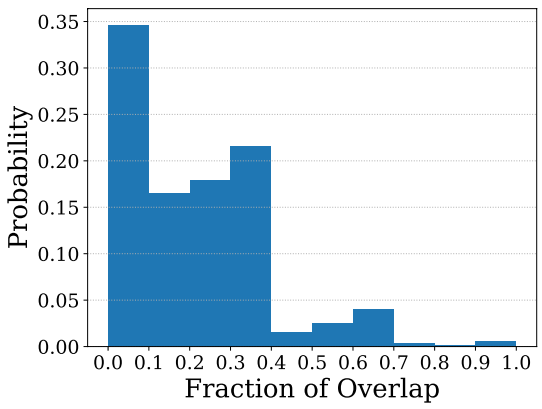


Fig. 3 Histogram of the fraction of ground-truth functional community overlaps across all of our Twitter datasets

We start by analysing the feasibility of finding identifiable structural patterns in ground-truth functional communities. Then, we evaluate structural community detection methods in the form of community scoring functions. Finally, we assess the goodness, robustness and sensitivity of those scoring functions when applied to Twitter functional ground-truth data.

4.1 Identifiable structural patterns

To provide preliminary evidence of distinctive structural patterns in the network of user interactions, we perform a comparative analysis of users in the ground-truth functional communities and randomly chosen connected nodes with the same path distribution (Yang and Leskovec 2015). If such distinctive connectivity patterns exist compared to randomly selected sets of connected nodes, we likely will be able to discover the functional communities based on their network connectivity.

We first define the sets of nodes that we will use in this analysis. For every ground-truth community C_i (of any type) in our datasets, we form a corresponding *non-community* $\tilde{C}_i$ from the user interactions network based on the following conditions:

1. Where possible, $\tilde{C}_i$ must be of the same size than C_i
2. Like every C_i, $\tilde{C}_i$ must also be connected
3. Where possible, users in $\tilde{C}_i$ must have the same distribution of shortest path distances of C_i

In microblogging, the first and third constraints are not easily satisfiable. We approach this problem by first computing the χ^2 distance (Pele and Werman 2010) between the shortest path length histograms of every C_i and of all potential candidates $\tilde{C}_i$. Then, for the first constraint, if it is not possible to find a non-community $\tilde{C}_i$ of the same size for a ground-truth community C_i, we select the closest candidate $\tilde{C}_i$ that has at least 75% of the size of C_i. Likewise, for the third constraint, if an exact match cannot be found, we instead select the closest candidate $\tilde{C}_i$ in descending order or randomly in case of multiple candidates with the same distribution.

We now define a set of structural properties that we will use to compare the structural patterns in the interactions network $G = (V, E)$ for both, ground-truth communities C_i and non-communities $\tilde{C}_i$. For this analysis, we will not consider the edge weights W.

Clustering Coefficient (CCF) measures how likely is a community to form a *small-world* cluster as defined in (Watts and Strogatz 1998).

Average Degree (AvgDeg) is the average node degree of the members of a community (Radicchi et al. 2004). AvgDeg is defined as: $2|E| / |V|$

Edge Density is the fraction of total edges possible in a community that are actually present (Radicchi et al. 2004). It is defined as: $2|E|/(|V|(|V|-1))$

Cohesiveness is the fraction of total edges possible in a community that are non-bridging (Leskovec et al. 2010). A non-bridge edge is such that when removed, the number of connected components is preserved. This measure captures the intuition of a well and evenly connected user community.

Table 2 Ratio between structural properties of ground-truth functional communities and randomly chosen nodes with similar shortest path distribution for the RTE2015 dataset

C. Type	CCF	AvgDeg	Density	Cohesiv
Cities	0.4034	1.0178	0.9169	0.5489
Countries	0.4365	0.9958	0.9510	0.6912
Hashtags	2.1117	1.2542	1.0885	2.0287
Mentions	3.7619	1.7942	1.3538	3.1683
Places	0.3981	0.9914	0.9329	0.4795
Quotes	2.3291	1.3907	1.1491	2.2839
Retweets	2.8460	1.6003	1.1834	2.6283
Urls	2.6746	1.2983	1.1495	2.4909
Average	1.8702	1.2928	1.0906	1.7900

In our analysis, each structural property p is computed for every C_i and $\tilde{C}_i$, and then the average ratio $r = p(C_i)/p(\tilde{C}_i)$ is computed for all community types in each dataset. If this average ratio $r > 1.0$, then we can assert that there is a measurable difference in the structural property p for C_i compared to $\tilde{C}_i$.

Example results for the RTE2015 dataset, which contains all the considered types of functional communities, are shown in Table 2. These results are similar in the other datasets. In this example, ground-truth functional communities have, in average, 87% higher clustering coefficient, 29% higher average degree, 9% higher edge density and 79% higher cohesiveness than their respective non-communities. This suggests that in fact the ground-truth has a distinctive structure compared to randomly chosen nodes in the same network.

In general, our results show that the ratio for each defined property is $r \geq 1.0$ in the majority of the ground-truth community types and datasets. The *mentions* community type excels in every dataset. A high ratio between communities C_i and non-communities $\tilde{C}_i$ can be observed, suggesting that functional communities with a third person as functional object are easier to discover than other types of social objects. Similar is the *hashtags* type, where it is only weak ($r < 1.0$) in the Pope2013 dataset. This is surprising because it suggests that users do not form strong communities around hashtags, despite the Pope Conclave event being highly susceptible to discussions around specific topics. The Ireland2017 dataset has very strong differentiable hashtags communities, e.g. $r > 30$ for CCF and $r > 15$ for cohesiveness. This is an interesting observation because the dataset was captured only using the location filter (Ireland), without forcing any particular topic. This outcome is likely because of the longer interaction times (A_i) as shown in Table 1.

For the *retweets* and *urls* functional types, most of our datasets also exhibit structurally distinguishable functional communities ($r > 1.0$). We note that Ireland2017 does not contain any retweets because Twitter does not deliver them for location-based capturing. This result is consistent with (Kwak et al. 2010), where retweeting and media links are regarded as core activities for news diffusion in Twitter.

The location-based functional types in general in our datasets contain few distinguishable communities, except for Ireland2017. This can be explained by the low signal of Tweets that actually contain useful location information (Hecht et al. 2011). The Ireland2017 dataset is captured using the location filter; Therefore, every Tweet in this dataset is embedded with location data. Nevertheless, the *countries* type was found to be distinctive enough (in most cases $r > 1.0$), suggesting that this abstraction is the most suitable for building functional communities based on location.

Finally, we also note that the *quotes* type is a recent functionality in Twitter, therefore is not present in datasets captured before 2015. Nonetheless, quotes have a strong differentiable structure ($r > 2.0$ for CCF and cohesiveness), suggesting that Twitter users seem to interact closely around quotes of interest to them.

4.2 Structural community scoring functions

The goal of community detection is to uncover sets of users in a network with a certain structural pattern. In this context, community scoring functions can be used to quantify how well a set of nodes fit to the desired structure. In this work, we consider thirteen commonly used community scoring functions pre-classified into four families for evaluation.

For a given set of nodes C, the scoring function $f(C)$ measures the quality of C as a structural community in an undirected network $G = (V, E)$. We define n_c and m_c as the number of nodes and edges in the set C, $n = |V|$ and $m = |E|$ as the number of nodes and edges in G, $d(v)$ as the degree of a node $v \in V$, and b_c as the number of edges on the boundary of C, i.e. edges that point outside of C. Using this notation, we now introduce the scoring functions under evaluation:

(*1*) Class → Only **Internal Connectivity**

- *Density*† fraction of total edges possible in C that are actually present. $f(C) = 2m_c/(n_c(n_c - 1))$
- *Edges Inside*† number of edges in C. $f(C) = m_c$
- *Avg. Degree*† average node degree of C. $f(C) = 2m_c/n_c$
- *Fraction over Median Degree*‡ **(FOMD)** fraction of nodes of C that have degree higher than d_m, where d_m is the median degree of all nodes $v \in V$. $f(C) = |\{u : u \in C, |\{(u, v) : v \in C\}| > d_m\}|/n_c$
- *Triangle Participation Ratio*‡ **(TPR)** fraction of nodes of C that belong to a triad. A triad is a set of three nodes that are fully connected to each other in C. $f(C) = |\{u : u \in C, u \text{ is in a triad}\}|/n_c$

(2) Class → Only *External Connectivity*

- *Expansion*† quantifies the number of edges per node in the boundary of C. $f(C) = b_c/n_c$
- *Cut Ratio*§ fraction of existing edges (out of all possible edges) leaving C. $f(C) = b_c/(n_c(n - n_c))$

(3) Class → *Internal* and *External Connectivity*

- *Conductance*†† fraction of total edge volume that is in the boundary of C. $f(C) = b_c/(2m_c + b_c)$
- *Normalised Cut*†† cost of cutting edges in C. $f(C) = b_c/(2m_c + b_c) + b_c/(2(m - m_c) + b_c)$
- *Maximum Out Degree Fraction*‡‡ **(ODF)** the maximum fraction of edges in C that point outside. $f(C) = \max_{u \in C}\left[|\{(u, v) \in E : v \notin C\}|/d(u)\right]$
- *Average-ODF*‡‡ is similar to Maximum-ODF but using the average measure instead of the maximum.
- *Flake-ODF*‡‡ fraction of nodes in C that have fewer edges pointing inside than outside of C. $f(C) = |\{u : u \in C, |\{(u, v) \in E : v \in C\}| < d(u)/2\}|/n_c$

(4) Class: *Network Model*

- *Modularity*§§ the difference between the number of edges m_c and its expectancy $E(m_c)$ in a random graph with identical degree sequence—a null model.

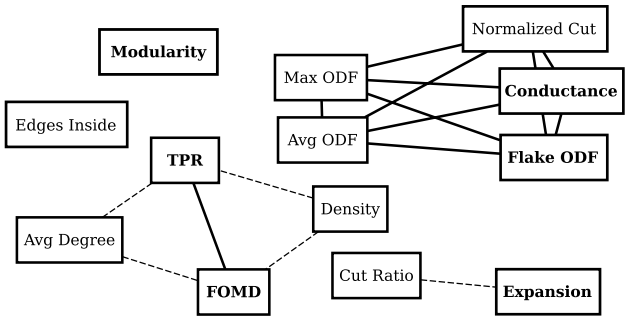


Fig. 4 Scoring functions clustered by correlation. Weak links ($\rho \geq 0.3$) are dashed and strong links ($\rho \geq 0.6$) are solid. All correlations were found significant with p values ≤ 0.05

The details for † can be found in Radicchi et al. 2004, for ‡ in Yang and Leskovec 2015, for § in Fortunato 2010, for †† in Shi and Malik 2000, for ‡‡ in Flake et al. 2000 and for §§ in Newman 2006.

Initially, we are interested in the relationship between these scores in our Twitter ground-truth data. To investigate the individual contribution of these scoring functions, we first compute each $f(C)$ for each of the 532,538 total ground-truth functional communities C we constructed. Then, we compute a correlation matrix based on the Pearson coefficient and filter it to unveil the correlation at different degrees between the scoring functions. As suggested in Evans 1996, we use $\rho \geq 0.3$ and $\rho \geq 0.6$ for weak and strong correlation, respectively, as thresholds. The result can be seen in Fig. 4, where the correlations were found significant with p values ≤ 0.05. With one exception, all of the scores grouped into four clusters, mirroring their pre-defined classes. The *Edges Inside* score remained isolated, even from its close relative *Avg. Degree*. This suggests that for the case of Twitter, considering only the size of the communities might be insufficient for their detection.

In general, this experiment suggests that despite having numerous structural definitions for communities, they mostly correlate in Twitter. For the remainder of this paper, we will focus on six representative scoring functions from the four classes (shown in bold in Fig. 4): FOMD, TPR, Cut Ratio, Conductance, Flake-ODF and Modularity.

4.3 Community detection goodness

We now evaluate the community scoring functions in terms of their quality to discover ground-truth functional communities in Twitter streams. In this experiment, we use goodness metrics that capture the notion that good communities should be compact, well connected and well isolated from the rest of the network. The difference between the goodness metrics and the scoring functions under study is that the first quantifies a desirable property of the communities, while the latter quantifies how community-like is a set of nodes. A community with high goodness does not imply a good scoring function value but a good community score should have a high goodness metric.

We consider four goodness metrics $g(C)$. Three of them—Density, Clustering Coefficient and Cohesiveness—were previously introduced as structural properties in Sect. 4.1. Therefore, a fourth new goodness metric is introduced below.

Separability captures the intuition that good communities should be well-distanced from each other (Fortunato 2010). This metric quantifies the ratio between the internal and external edges.

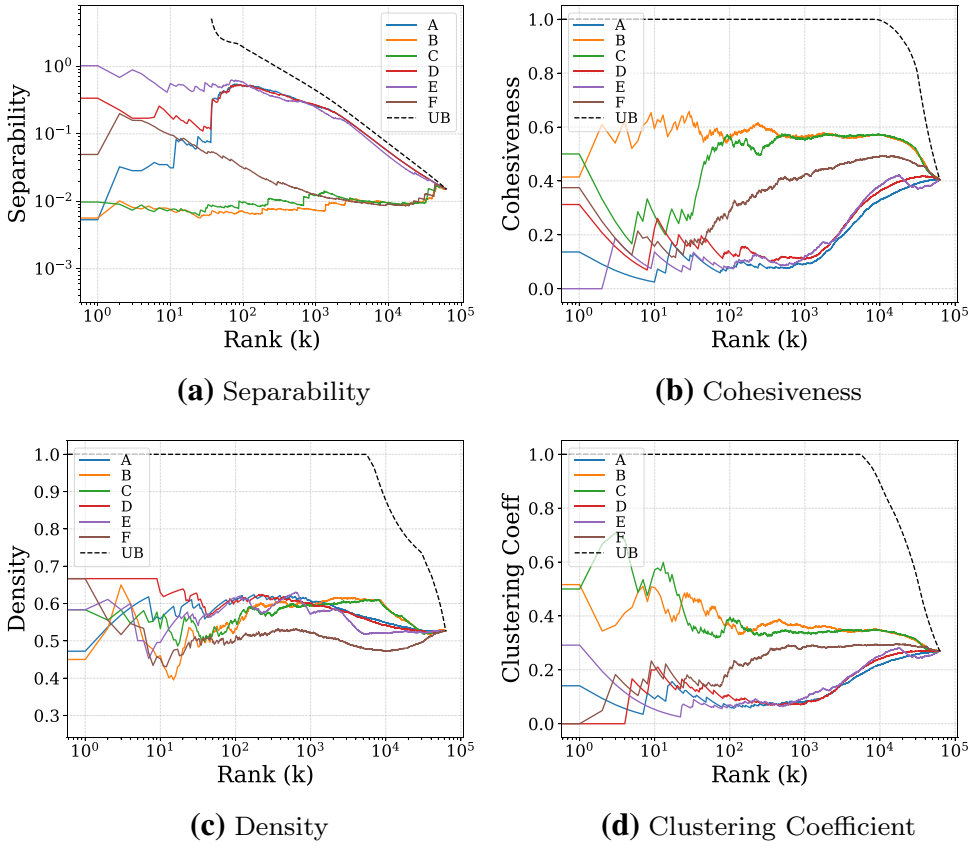
We set up the goodness experiment as follows. For each dataset and community type, we rank our ground-truth functional communities C_i using the six selected scoring functions $f(C_i)$ in descending order. Then, we measure the cumulative moving average (CMA) of each goodness metric $g(C_i)$ for the top-k ground-truth communities under the order induced by $f(C_i)$. A perfect scoring function should rank the ground-truth communities in the same descending order as the goodness metrics, and therefore the CMA should decrease monotonically along k. Conversely, a poor community scoring function would produce a k-dependent constant CMA.

The results were similar across all of our datasets. For the remainder of this Section, we report aggregated results using all the community types of Ireland2017 as an example. Figure 5 shows the four ranked goodness metrics for this representative example, where the CMA of the six representative scoring functions ranked by the four goodness metrics can be seen. An additional upper bound curve (e.g. the CMA of separability ranked by separability) is also provided for reference.

We observe that Cut Ratio (A), Conductance (D) and Flake-ODF (E) have a near perfect fit in separability, while FOMD (B) and TPR (C) show instead an inverse ordering, suggesting that the latter two prefer more dense communities. If the analyst desires denser communities regardless of separation, FOMD and TPR should be preferred. In contrast, for Cohesiveness, Density and Clustering Coefficient, FOMD (B) and TPR (C) prevail with good performance. These not only prefer denser but also cohesive and tight communities.

Modularity (F) performs relatively well in Cohesiveness. However, Cut Ratio (A), Conductance (D) and Flake-ODF (E) exhibit inverse ordering. This observation suggests that these scores prefer sparse communities, revealing their inability to properly capture cohesive groups in the Twitter microblogging scenario.

A reverse ordering of Modularity (F) for the Density goodness metric can be also observed. This is a



(a) Separability **(b)** Cohesiveness

(c) Density **(d)** Clustering Coefficient

Fig. 5 Ranked Scoring Functions by CMA for all community types in the IRELAND2017 dataset. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake-ODF (E), Modularity (F) and their Upper Bound (UB)

manifestation of the well-known resolution limit of the Modularity score (Fortunato and Barthélemy 2007) that becomes evident in our ground-truth functional communities. Modularity also exhibits a near-constant ranking for Clustering Coefficient, suggesting that this score does not prefer nor reject well-packed communities.

To complement our results, we also study the ability of the scoring functions to rank the ground-truth communities using the goodness metrics. For each goodness metric $g(C)$ and scoring function $f(C)$ in all of our datasets, we observe the rank of each score in comparison with the other scoring functions at every rank k. For example, in Fig. 5 for Clustering Coefficient at $k = 10^2$, the scores are ranked as: 1st TPR (C), 2nd FOMD (B), 3th Modularity (F), 4th Conductance (D), 5th Flake-ODF (E) and 6th Cut Ratio (A). Therefore, for every k, we rank and aggregate the six scores using the Borda voting method (Saari 2012) to obtain a unified ranking that quantifies the ability of each scoring function to find *good communities*. The results are in Table 3, where ranks ≈ 1.0 indicate scoring functions adequate for each goodness criteria.

Overall, to identify more clustered, dense and cohesive communities in Twitter, FOMD and TPR are the better choices. If dense but more separated communities are preferred, then Conductance is more adequate.

4.4 Community detection robustness

We now investigate the robustness and sensitivity of the structural community scoring functions in the presence of different random perturbations to the ground-truth functional communities. A good scoring function should be stable under small perturbations and reduce its performance under strong disturbance. The perturbation strategies we consider (Yang and Leskovec 2015) are below.

- *NodeSwap* simulates the effect of community users diffusing from C through the network. First, a random edge

Table 3 Aggregated scoring ranking by goodness metrics using the Borda voting method for all ground-truth datasets

Score	CCF	Cohesiv	Density	Separ
Cut ratio	5.4589	5.7208	3.3097	2.2203
FOMD	2.0974	**1.0014**	**1.0001**	5.0531
TPR	**1.1422**	2.2890	3.4189	5.6131
Conductance	4.1373	3.9755	2.8899	**1.3169**
Flake-ODF	5.2929	5.2416	4.8373	2.4659
Modularity	2.8712	3.0755	5.5441	4.3307

Best ranked scoring functions for each goodness metric are in bold

$(u, v), u \in C, v \notin C$ is chosen, and then the nodes u and v are swapped. This causes u to abandon C and v to join.

- *Random* perturbs communities by swapping a random member $u \in C$ with a random non-member $v \notin C$. Similarly to NodeSwap, Random does not change the size of C but may disconnect it.
- *Expand* increases the size of communities by choosing random non-members $v \notin C$ that are connected to members $u \in C$, and incorporating them into C. This action decreases the quality of the community.
- *Shrink* decreases the size of communities by choosing random boundary edges $(u, v), u \in C, v \notin C$ and removing the user u from C. Similarly to Expand, this perturbation preserves the connectedness.

The above strategies can be controlled using an intensity parameter p that specifies the number of times ($p|C|$) the perturbation is applied to a community C.

To quantify the impact of applying any perturbation strategy h to a given ground-truth functional community C, let's consider $h(C, p)$ the perturbed version of C under perturbation h with intensity p. Then, we measure the Z-score (units of standard deviation) of the difference between the score $f(C)$ of the unperturbed community C and the score $f(h(C, p))$, as seen below.

$$Z(f, h, p) = \frac{E\left[f(C_i) - f(h(C_i, p))\right]}{\sqrt{\mathrm{Var}\left[f(h(C_i, p))\right]}} \tag{4}$$

In Eq. 4, $E[\cdot]$ is the expectancy operator (the mean) and $Var[\cdot]$ is the variance operator, both applied over all the ground-truth communities C_i. We note that the sign of TPR, FOMD and Modularity needs to be inverted to ensure that all scores have the same interpretation, i.e. higher is better. Due to the random nature of the proposed perturbations, we repeat the experiment 20 times and average the resulting Z-scores.

The experiment is as follows. We vary the perturbation intensities $p \in [0.01, 0.60]$—e.g. in the NodeSwap strategy, this means exchanging between 1 and 60% of the members of a community—and observe the averaged Z-score across all ground-truth functional communities for each community type and dataset. The results were similar in all of our datasets. For the remainder of this Section, we report aggregated results using all the community types of RTE2015 as an example. Figure 6 shows the Z-scores for each perturbation strategy and scoring function for the representative example.

The TPR and FOMD scores perform the best in the NodeSwap experiment, followed by Conductance and Flake-ODF. In contrast, Modularity and Cut Ratio do not degrade gracefully when we increase the perturbation, revealing their inability to deal with noisy data in Twitter. Similar observations can be made for the Random strategy, where Flake-ODF takes the lead and TPR/FOMD fall behind. Cut Ratio performs the worst in the presence of strong noise.

The Expand and Shrink strategies also reveal TPR and FOMD as robust scores for Twitter functional communities, and Flake-ODF and Cut Ratio being ineffective in Expand and Shrink, respectively. Modularity has consistent good performance in small intensities for Expand and large for Shrink but degrades with larger expansions and smaller reductions. This is again evidence of its resolution limit (Fortunato and Barthélemy 2007).

In this experiment, TPR and FOMD proved to be robust community scoring functions for Twitter interaction streams, while Modularity and Cut Ratio proved weaker in the same context. Alternatively, Conductance and Flake-ODF are also good; however, the latter is not stable in expanding and shrinking communities.

Finally, we explore how sensitive are the scoring functions in terms of small and large perturbations. We measure the change of Z-score between a small ($p = 0.04$) and a large ($p = 0.20$) perturbation, giving preference to scoring functions that quickly degrade in the presence of strong perturbations. We averaged the difference $Z(f, h, 0.20) - Z(f, h, 0.04)$ across all our ground-truth functional communities and the results can be seen in Table 4. Large differences indicate that the community scoring function is both robust and sensitive.

In general, FOMD stands as the most robust and sensible score in this experiment for all the perturbation strategies. Conductance is a close second best for NodeSwap, while TPR is in all the others. On the other hand, Modularity performs the worst under every perturbation except Shrink, where Cut Ratio is placed last.

5 Evaluating communities: dynamic scenario

In this Section, research questions (RQ4) and (RQ5) are addressed: how can user activity hotspots be identified inside the ground-truth functional communities? and how do existing structural community definitions perform when applied to these activity hotspots?

User activity hotspots and an approach to extract and smooth user activity from ground-truth communities were introduced in Sect. 3.3. We hypothesise that given the fast paced and sparse nature of microblogging user interactions, user communities based on social functions are of better structural quality when focusing on particular portions using their activity over time. This improvement also translates into better performance for the community detection task using current state-of-the-art community detection approaches.

We first investigate the identification of hotspots in the defined ground-truth functional communities and then

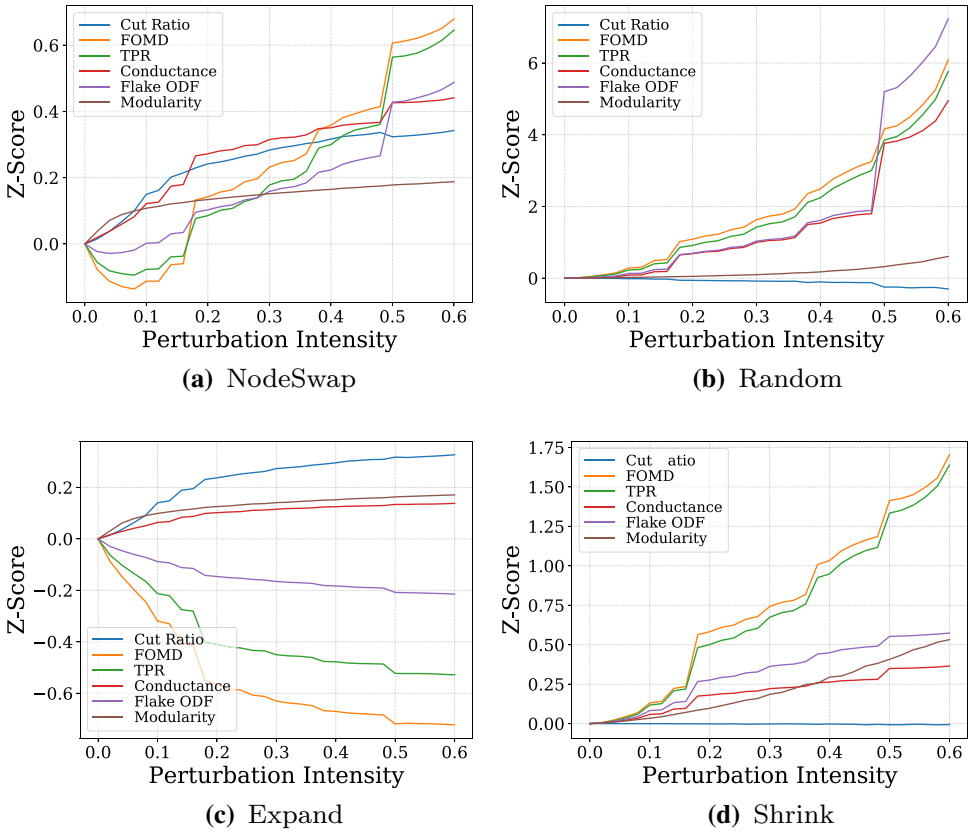


Fig. 6 *Z*-scores of intensities for perturbation strategies applied to all community types in the RTE2015 dataset

Table 4 Average absolute increment in *Z*-score between small and large community perturbations

Score	N. Swap	Rndm	Expnd	Shrnk
Cut ratio	0.1050	0.0588	0.1114	0.0020
FOMD	**0.4800**	**1.1651**	**0.3058**	**0.7006**
TPR	0.2675	0.8517	0.1665	0.5194
Conductance	0.3620	0.8128	0.0918	0.2331
Flake-ODF	0.2854	0.8384	0.1214	0.3254
Modularity	0.0699	0.0275	0.0675	0.0501

Best are in bold

evaluate existing structural scoring functions and their goodness for the found activity hotspots.

5.1 Identifying activity hotspots in communities

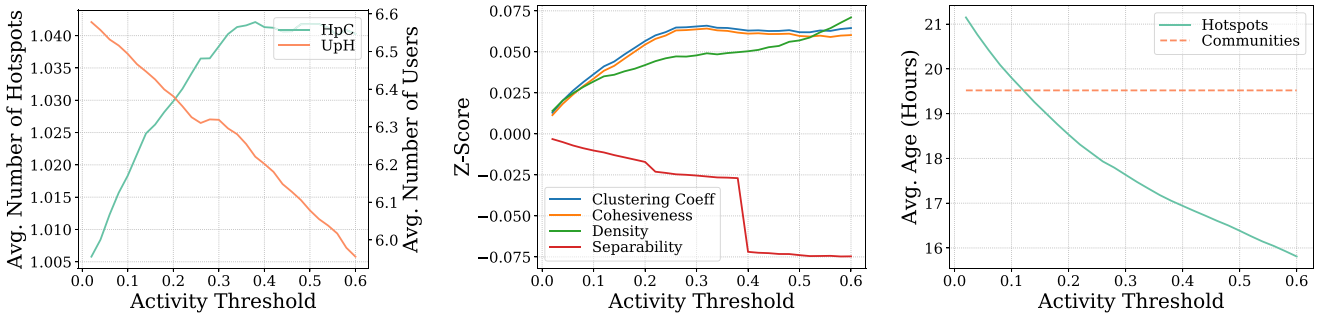
We aim to identify user activity hotspots H in ground-truth communities C from microblogging. For this, we propose to extract the user activity from each ground-truth functional community C and apply an exponential decay function for smoothing as described in Sect. 3.3. To obtain a comparable range of values across all the ground-truth communities under study, we apply min–max normalisation over the smoothed activity.

We define a user activity threshold $\alpha \in [0, 1]$ that we will use to find the starting and ending points in time for user activity hotspots H inside each ground-truth community C in our datasets. Each time the normalised user activity of C rises above α, we record the observed time as the starting time boundary of a new hotspot. When the activity falls below α, we record the observed time as the ending boundary of the same hotspot.

Choosing the threshold α will depend on the communities that the analyst is interested on discovering. However, we propose a systematic method in this paper for finding a reasonable threshold for each of our ground-truth datasets to further investigate the performance of community scoring functions in Sect. 5.2.

The method is as follows. First, we define two quantitative metrics for activity hotspots that measure the characteristic of a set of generated hotspots given a particular α. These metrics are defined below.

- *Avg. Hotspots per Community (HpC)* measures, in average, how many activity hotspots are generated per ground-truth functional community. The higher is HpC, the more fragmented the original ground-truth communities will become.
- *Avg. Users per Hotspot (UpH)* measures, in average, how many users are assigned to each generated hotspot. The



(a) Hotspots Metrics **(b)** Hotspots Goodness **(c)** Hotspots Ages

Fig. 7 Activity thresholds α and their effect over the activity hotspots metrics (HpC and UpH), hotspots goodness metrics and hotspots ages for all the community types combined of the Pope2013-Spl dataset

Table 5 Selected activity hotspots thresholds for all of the ground-truth datasets

Dataset	Threshold (α)	Criterion	Value	CCF	Cohesiveness	Density	Separability
Pope2003	0.24	HpC	1.2043	0.0360	0.0407	0.0129	− 0.2165
Pope2003-Spl	0.28	UpH	6.3198	0.0650	0.0633	0.0471	− 0.0250
WorldCup2014	0.10	HpC	2.2583	0.0650	0.0825	− 0.0913	− 0.2370
RTE2015	0.04	HpC	6.7229	0.4810	0.4836	0.2924	− 0.9583
Ireland2017	0.04	HpC	7.2813	0.7003	0.6701	0.3675	− 0.2866

For each selected threshold, the selection criterion—*hotspots per community* (Hpc) or *user per hotspots* (UpH)—and its value at the selected threshold are shown. Furthermore, the *Z*-score of each goodness metric at the same selected threshold is also reported

higher is UpH, the more concentrated the generated hotspots will become.

A good activity threshold α should maximise both, HpC and UpH. Therefore, we will use these metrics as criteria for selecting α within a range of candidates. We setup the experiment by varying α in the range $[0.02, 0.60]$ in increments of 0.02, and measuring HpC and UpC on each set of generated hotspots. We perform this for every community type in all of our datasets. An example result of the above for all the community types of the Pope2013-Spl dataset can be seen in Fig. 7a.

To select an activity threshold α, we will prefer the metric that exhibits the higher statistical coefficient of variation as criterion. Once a criterion metric is selected, we apply a simple peak finding algorithm[1] to identify the global maxima for that criterion.

A summary with all the activity thresholds α selected using this method for all of the community types contained in each of our ground-truth datasets is found in Table 5. We observe the selected criterion metric based on the highest variation and the value of this selected metric measured at the chosen threshold.

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html.

Note that all the selected thresholds are $\alpha < 0.30$, suggesting that no more than 30% of the relative user activity is required as trigger for forming reasonable hotspots in our ground-truth communities. Also, the dominant criterion is HpC, suggesting that the average number of hotspots per community is preferred over the average amount of users in them for selecting the activity threshold. In Sect. 5.2, we further demonstrate the quality of the selected thresholds.

To complement our study, we also investigated the effects of the activation threshold α on the temporal properties of the generated activity hotspots. In this experiment, the age of an activity hotspot is measured as the difference between the start and end times of the hotspot timespan for every activation threshold α in [0.02, 0.60]. The results are in Fig. 7c for all the community types of the Pope2013-Spl dataset (as previously), where the average age of the original ground-truth communities is also provided as reference.

In general, the activity hotspots have much less age than the communities they originate from. The sole exception is in the Pope2013-Spl dataset (in the figure), where at $\alpha < 0.15$ the hotspots live longer than the average ground-truth communities. This is due to this particular dataset being randomly sampled, and therefore many ground-truth communities are very short lived. In this case, the identification of user activity hotspots can potentially lead to more meaningful groups of users, i.e. that are engaged for longer time. The

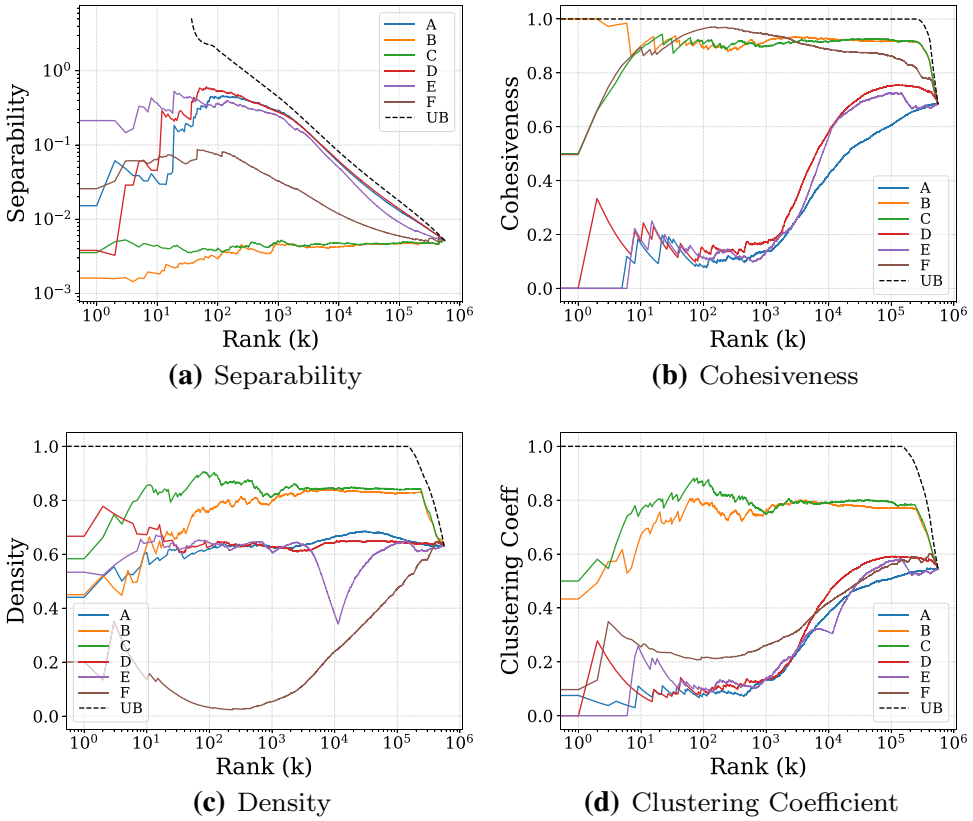


Fig. 8 Ranked Scoring Functions by CMA for all hotspots in all community types in the IRELAND2017 dataset. Scores: Cut Ratio (A), FOMD (B), TPR (C), Conductance (D), Flake-ODF (E), Modularity (F) and their Upper Bound (UB)

hotspots ages also decrease when the activation threshold α is increased because, at higher values of α, the hotspots contain less users and therefore less activity.

5.2 Improving community detection using hotspots

We now investigate the goodness metrics from Sect. 4 applied to the newly formed activity hotspots. Our goal is to demonstrate that when considering user activity for the identification of activity hotspots, the resulting sub-communities are of better quality when compared to the originating communities. We show evidence of this by measuring the goodness of the activity hotspots generated from our ground-truth functional communities and comparing it to the same goodness metrics previously discussed in the static scenario evaluation.

In this experiment, we measure the Z-score similarly as in Eq. 4 for the differences in goodness metrics between all the original ground-truth functional communities of each dataset and the newly formed activity hotspots using the selected thresholds α in Table 5. An example result for the POPE2013-SPL dataset can be seen in Fig. 7b. Note how the goodness metrics in this example are nearly maximised at the selected threshold $\alpha = 0.28$, with the exception of separability. In fact, this result can be observed in all of our datasets, where separability never improved when considering activity hotspots. This suggests that while the activity hotspots improve in terms of Clustering Coefficient, Density and Cohesiveness, they do not become more separable. This outcome is not surprising because all the activity hotspots from a given community can only be formed with the same users of that community.

To complement our results, we also perform the ranked goodness experiment from Sect. 4 using the constructed activity hotspots instead of the original ground-truth functional communities. The results of the experiment using all the community types for the IRELAND2017 dataset can be seen in Fig. 8. We chose this particular example because it can be directly compared to the ranked goodness results shown in Fig. 5 for the static scenario evaluation. Again, these results are consistent across our datasets.

It can be observed that most of the scoring functions drastically improve their alignment to the goodness metrics when considering the proposed activity hotspots. In particular, FOMD (B) and TPR(C) now achieve nearly perfect fit with the upper bounds (UB). A remarkable improvement is Modularity (F) for the Cohesiveness goodness metric, as now it is able to attain almost the same performance than FOMD and TPR. As previously discussed, using activity hotspots does not significantly impact the separability goodness metric.

Overall, the activity hotspots are able to improve the performance of structural scoring functions in terms of Cohesiveness, Density and Clustering Coefficient, but not separability, which remains mostly unaffected.

6 Conclusions and future directions

In this work, we address the problem of evaluating community detection in the context of microblogging services, represented by Twitter. For this, we adopted two interpretations of community: a *functional* definition (used as ground-truth) based on user-labelled social functions, and a *structural* definition based on the connectivity patterns in a network.

We proposed several research questions with the goal of evaluating the adoption of user interactions networks and the construction of explicitly labelled ground-truth functional communities using varied social objects present in Twitter streams. Additionally, we thoroughly evaluated a set of structural community scoring functions from different classes using our ground-truth in both, a static and dynamic scenario.

For the static scenario, we conclude that scoring functions based on internal connectivity such as the TPR, FOMD and Conductance work best for Twitter, proving to be robust and sensitive. Conversely, the popular Modularity score is limited and unfit due to the sparse and noisy characteristics of microblogging.

For the dynamic scenario, we conclude that the construction of activity hotspots from ground-truth functional communities further improves the ability of scoring functions such as TPR, FOMD and Conductance to discover community in Twitter. Moreover, the Modularity score was shown to be less limited in this context.

We believe that more research is required to better understand the nature of microblogging and the community detection task for it. For example, we considered native hashtags as the social function for topics; however, other models can be used, e.g. named entities, bag-of-words or TF-IDF. Also, the construction and exploitation of user activity hotspots can be further improved. For example, the λ parameter and the α threshold can be learned instead from empirical data using machine learning techniques. Furthermore, the current findings in this work such as the activity threshold can be potentially applied to the design of an automatic windowing approach for real-time community detection using stream processing.

Note Twitter policies prevent us to share our datasets, however we make available, on request, our framework based on the SNAP engine (Leskovec and Krevl 2015) used to obtain the reported results.

References

Akbari M, Chua TS (2017) Leveraging behavioral factorization and prior knowledge for community discovery and profiling. In: Proceedings of the tenth ACM international conference on web search and data mining, ACM, New York, NY, USA, WSDM '17, pp 71–79. https://doi.org/10.1145/3018661.3018693

Amor BR, Vuik SI, Callahan R, Darzi A, Yaliraki SN, Barahona M (2016) Community detection and role identification in directed networks: understanding the twitter network of the care data debate. In: Dynamic networks and cyber-security, pp 111–136. http://spiral.imperial.ac.uk/handle/10044/1/32881

Aslak U, Rosvall M, Lehmann S (2018) Constrained information flows in temporal networks reveal intermittent communities. Phys Rev E 97(6):062312. https://doi.org/10.1103/PhysRevE.97.062312

Aslam S (2018) Twitter by the numbers (2018): stats, demographics & fun facts. Retrieved 9 Jan 2018, from https://www.omnicoreagency.com/twitter-statistics/

Bakillah M, Li RY, Liang SHL (2015) Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. Int J Geogr Inf Sci 29(2):258–279. https://doi.org/10.1080/13658816.2014.964247

Cao N, Lu L, Lin YR, Wang F, Wen Z (2015) SocialHelix: visual analysis of sentiment divergence in social media. J Vis 18(2):221–235. https://doi.org/10.1007/s12650-014-0246-x

Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th international conference on World Wide Web, ACM, New York, NY, USA, WWW '09, pp 721–730. https://doi.org/10.1145/1526709.1526806

Darmon D, Omodei E, Garland J (2015) Followers are not enough: a multifaceted approach to community detection in online social networks. PloS One 10(8):e0134860. https://doi.org/10.1371/journal.pone.0134860

Engeström J (2005) Why some social network services work and others don't—Or: the case for object-centered sociality. Retrieved 10 May 2018, from Zengestrom. http://www.zengestrom.com/blog/2005/04/why-some-social-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html

Evans JD (1996) Straightforward statistics for the behavioral sciences. Brooks/Cole Publishing Company, Baltimore

Flake GW, Lawrence S, Giles CL (2000) Efficient identification of web communities. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '00, pp 150–160. https://doi.org/10.1145/347090.347121

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci 104(1):36–41. https://doi.org/10.1073/pnas.0605965104

Gupta A, Joshi A, Kumaraguru P (2012) Identifying and characterizing user communities on Twitter during crisis events. In: Proceedings of the 2012 workshop on DUBMMSM, ACM, New York, NY,

USA, DUBMMSM '12, pp 23–26. https://doi.org/10.1145/2390131.2390142
Hecht B, Hong L, Suh B, Chi EH (2011) Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, NY, USA, CHI '11, pp 237–246. https://doi.org/10.1145/1978942.1978976
Hromic H (2019) Methods for defining dynamic online communities and community detection in fast-paced social media streams. Thesis, NUI Galway, https://aran.library.nuigalway.ie/handle/10379/15146
Hromic H, Hayes C (2014) Constructing Twitter datasets using signals for event detection evaluation. In: Synergies of case-based reasoning and data mining workshop
Hromic H, Hayes C (2018) Characterising and evaluating online communities from live microblogging user interactions. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 21–24. https://doi.org/10.1109/ASONAM.2018.8508392
Hromic H, Prangnawarat N, Hulpuş I, Karnstedt M, Hayes C (2015) Graph-based methods for clustering topics of interest in Twitter. In: Cimiano P, Frasincar F, Houben GJ, Schwabe D (eds) Engineering the web in the big data era. Lecture notes in computer science. Springer, Berlin, pp 701–704
Hromic H, Barraza-Urbina A, Hayes C, Cantle N (2017) Mining TV twitter networks for adaptive content navigation and community awareness. Expert Update 17(1)
Java A, Song X, Finin T, Tseng B (2007) Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, ACM, New York, NY, USA, WebKDD/SNA-KDD '07, pp 56–65. https://doi.org/10.1145/1348549.1348556
Kietzmann JH, Hermkens K, McCarthy IP, Silvestre BS (2011) Social media? Get serious! Understanding the functional building blocks of social media. Bus Horiz 54(3):241–251. https://doi.org/10.1016/j.bushor.2011.01.005
Kumar R, Novak J, Tomkins A (2010) Structure and evolution of online social networks. In: Yu PS, Han J, Faloutsos C (eds) Link mining: models, algorithms, and applications. Springer, New York, pp 337–357. https://doi.org/10.1007/978-1-4419-6515-8_13
Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, ACM, New York, NY, USA, WWW '10, pp 591–600. https://doi.org/10.1145/1772690.1772751
Leskovec J, Krevl A (2015) SNAP datasets: stanford large network dataset collection
Leskovec J, Lang KJ, Mahoney M (2010) Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th international conference on World wide web, ACM, pp 631–640
Lu X, Brelsford C (2014) Network structure and community evolution on Twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami. Sci Rep 4:6773. https://doi.org/10.1038/srep06773
Mcauley J, Leskovec J (2014) Discovering social circles in ego networks. ACM Trans Knowl Discov Data 8(1):4. https://doi.org/10.1145/2556612
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27(1):415–444. https://doi.org/10.1146/annurev.soc.27.1.415
Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. AAAI Press, Palo Alto
Myers SA, Leskovec J (2014) The bursty dynamics of the Twitter information network. In: Proceedings of the 23rd international conference on World Wide Web, ACM, New York, NY, USA, WWW '14, pp 913–924. https://doi.org/10.1145/2566486.2568043
Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103
Newman MEJ, Park J (2003) Why social networks are different from other types of networks. Phys Rev E 68(3):036122. https://doi.org/10.1103/PhysRevE.68.036122
Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818. https://doi.org/10.1038/nature03607
Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446(7136):664–667. https://doi.org/10.1038/nature05670
Papadopoulos S, Kompatsiaris Y, Vakali A, Spyridonos P (2011) Community detection in social media. Data Min Knowl Discov 24(3):515–554. https://doi.org/10.1007/s10618-011-0224-z
Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. Sci Adv 3(5):e1602548. https://doi.org/10.1126/sciadv.1602548
Pele O, Werman M (2010) The quadratic-chi histogram distance family. In: Computer vision—ECCV 2010, Springer, Berlin, Heidelberg, Lecture notes in computer science, pp 749–762.https://doi.org/10.1007/978-3-642-15552-9_54
Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci USA 101(9):2658–2663. https://doi.org/10.1073/pnas.0400054101
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123. https://doi.org/10.1073/pnas.0706851105
Saari DG (2012) Geometry of voting. Springer, Berlin
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, ACM, New York, NY, USA, WWW '10, pp 851–860. https://doi.org/10.1145/1772690.1772777
Shamma DA, Kennedy L, Churchill EF (2009) Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on social media, ACM, New York, NY, USA, WSM '09, pp 3–10. https://doi.org/10.1145/1631144.1631148
Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905. https://doi.org/10.1109/34.868688
Sundaram H, Lin Y, Choudhury MD, Kelliher A (2012) Understanding community dynamics in online social networks: a multidisciplinary review. IEEE Signal Process Mag 29(2):33–40. https://doi.org/10.1109/MSP.2011.943583
Tang L, Liu H (2010) Community detection and mining in social media. Synth Lect Data Min Knowl Discov 2(1):1–137. https://doi.org/10.2200/S00298ED1V01Y201009DMK003
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440. https://doi.org/10.1038/30918
Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowl Inf Syst 42(1):181–213. https://doi.org/10.1007/s10115-013-0693-z
Yang Y, Lan C, Li X, Luo B, Huan J (2014) Automatic social circle detection using multi-view clustering. In: Proceedings of the 23rd ACM CIKM, ACM, New York, NY, USA, CIKM '14, pp 1019–1028. https://doi.org/10.1145/2661829.2661973
Zhou W, Jin H, Liu Y (2012) Community discovery and profiling with social messages. In: Proceedings of the 18th ACM SIGKDD, ACM, pp 388–396. https://doi.org/10.1145/2339530.2339593